DOCUMENT RESUME

ED 406 214 SE 059 951

AUTHOR DeMars, Christine
TITLE Physics or Biology? Geometry or Algebra? Gender and Content Area Interactions on a High School Proficiency Test.
PUB DATE Mar 97
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Achievement Tests; Grade 11; High Schools; Knowledge Level; Mathematical Aptitude; Mathematics Curriculum; *Mathematics Tests; Science Curriculum; *Science Tests; *Sex Differences; Units of Study
IDENTIFIERS Michigan

ABSTRACT

The degree of gender differences in mathematics and science appears to vary with the content subdomain. Differences also appear to be greater on items assessing content knowledge than with items measuring reasoning about scientific processes. Many studies of gender differences have involved fairly select populations. This study focuses instead on a broad population, representing a random sampling of almost all high school juniors in Michigan. The focus of this study is the pilot results from the science and math portions of the Michigan High School Proficiency Test (HSPT), a diploma endorsement test that includes both constructed response and multiple-choice items. Two of the pilot forms of the mathematics and science sections of the HSPT were examined for gender by content scale interactions. Other studies had found gender differences to be greater in geometry (compared with algebra) and physical and earth sciences (compared with life sciences and process-oriented science items). These findings were generally not replicated with the HSPT (except among students above the 95th percentile on the mathematics test). Correlations among the subscales were similar for boys and girls as were the standard errors of measurement for each scale. (Author/PVD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 406 214

Physics or Biology? Geometry or Algebra?

Gender and Content Area Interactions on a High School Proficiency Test

Christine DeMars

Measurement and Quantitative Methods

Michigan State University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
C. DeMars
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A poster session presented at the Annual Meeting of the American Educational Research Association, March 24-28, 1997, Chicago.

BEST COPY AVAILABLE

Two of the pilot forms of the mathematics and science sections of the Michigan High School Proficiency Test (HSPT) were examined for gender by content scale interactions. Other studies had found gender differences to be greater on geometry (compared to algebra) and physical and earth sciences (compared to life sciences and process-oriented science items). These findings were generally not replicated on the HSPT (except among the students above the 95th percentile on the mathematics test). Correlations among the subscales were similar for boys and girls, as were the standard errors of measurement for each scale.

The degree of gender differences in mathematics and science appears to vary with the content subdomain. In science, the gender differences tend to be greatest in physics and least in biology (Becker, 1989; Comber & Keeves, 1973; Erickson & Erickson, 1984; Stanley, 1987). Differences also tend to be greater on items assessing content knowledge compared to items measuring reasoning about scientific processes (Erickson & Erickson, 1984; Linn, De Benedictis, Delucchi, Harris, & Stage, 1987; Linn & Hyde, 1989). In math, the findings are more mixed, but among high school students (and one sample of 8th graders) the males tend to do relatively better on geometry items and applied items and females tend to do relatively better on algebra items (Doolittle & Cleary, 1987; Harris & Carlton, 1993; Ryan & Fan, 1996).

The focus of this study is the pilot results on the science and math portions of the Michigan High School Proficiency Test (HSPT), a diploma endorsement test which includes both constructed response and multiple choice items. The content of this test is above the "minimal competency" level of some state tests, but is lower than the level of some college entrance exams: the objectives cover competencies Michigan students should have had the opportunity to achieve by the end of tenth grade.

Many of the studies of gender differences in mathematics have involved fairly select populations (Becker, 1990; Doolittle & Cleary, 1987; Harris & Carlton, 1993). Using meta-analytic techniques, Feingold (1992) concluded the gender gap in quantitative abilities is larger at the upper end of the distribution because of the greater variance in males' scores. Though others (Hedges & Friedman, 1993; Katzman & Alliger, 1992) have suggested alternative methods which result in less extreme differences than Feingold's, the finding of larger differences among high-ability subjects remains. This study instead focuses on a broad population; almost all high school juniors in Michigan will take the HSPT, and the pilot schools were chosen to be representative of schools in the state. Gender differences often vary in different ranges of the ability distribution, so

this research reveals more about "typical" high school students. In addition to including a wide range of ability levels, the HSPT has the advantage of testing the same students on all subareas within a content area; on tests such as the AP science exams where different students choose to take different exams, the students taking the physics test, for example, may not be from the same areas of each gender's ability distribution as the students taking the biology test.

Method/Data

Subjects. A stratified, random sample of schools was selected to participate in the pilot. All regular education students in the selected schools who were present on the day of testing were to be tested. No school was asked to participate in the pilot test of more than one content area. 102 schools participated in the math pilot, and 99 other schools participated in the science pilot. Multiple forms of each test were administered, and two forms in each area were selected for this study.

Instrument. The Michigan High School Proficiency Test (HSPT) has four components: math, science, reading, and writing. Beginning with the graduating class of 1997, students who pass the appropriate sections will receive diploma endorsements in math, science, and language arts (which will require a passing score on both reading and writing). The test is *not* designed as a minimum competency exam; it is intended to reflect high school level (through the end of the sophomore year) skills.

The HSPT mathematics test contains 40 multiple choice items. Content areas tested include numbers/number systems, algebraic reasoning, geometry, and data interpretation. Students are allowed to use a calculator. The HSPT science test has 42 multiple choice items. Items are distributed in five categories: Using Life Science, Using Earth Science, Using Physical Science, Constructing Knowledge, and Reflecting on Knowledge. The number of items in each

category varies from form to form. Item specifications for the constructing and reflecting items did not dictate a specific content area; item writers could use whatever content, or mix of content, would best test the objective.

Results

As noted, only two forms of each test were analyzed, and each was analyzed separately (rather than including form as a factor within a single design). This is because the forms have not been equated on a subscale level, and on the science test, the number of points on each scale varies from form to form. One form for each subject area (designated form A here) was the form which was administered in Spring 96. The first operational tests were chosen to be analyzed here because these will be the first operational results available for comparison to the pilot results. Differences from pilot to operational forms may influence the interpretation of results of the remaining pilot forms, which are to be used as operational tests in later administrations. Form B in each subject area was selected randomly. These forms were used to check the consistency of the findings from the A forms.

Descriptive statistics for the content scales appear in Tables 1 and 2. Scores are given in proportion correct, rather than raw score points, to make it easier to compare different scales (in science, the number of points on each scale is different). The differences are also calculated in standard deviation units (the difference between the female mean and the male mean divided by the square root of the pooled variance--positive differences indicate females scored higher).

In math, on form A, males scored *slightly* higher (about 1 percentage point) than females on every scale except geometry. On form B, males scored slightly higher on every scale, with the largest differences in geometry and data analysis. In science, on form A females scored higher than males on reflecting on scientific knowledge and on life science; males scored higher on the other scales, with the smallest gender difference on constructing knowledge. On form B, males

scored higher on all scales, and the differences were smallest on the constructing and life science scales.

Table 1
Means and Standard Deviations for Math Scales

Form A		Gender	N	Mean	SD	Difference in standard dev.
	Numbers	Female	603	0.597	0.211	-0.05
		Male	572	0.608	0.222	
	Data Analysis	F	603	0.586	0.224	-0.04
		M	572	0.596	0.229	
	Geometry	F	603	0.635	0.201	0.10
		M	572	0.615	0.216	
	Algebra	F	603	0.592	0.241	-0.05
		M	572	0.604	0.240	
Form B						
	Numbers	F	694	0.591	0.228	-0.05
		M	652	0.604	0.249	
	Data Analysis	F	694	0.550	0.218	-0.10
		M	652	0.573	0.228	
	Geometry	F	694	0.572	0.242	-0.15
		M	652	0.609	0.256	
	Algebra	F	694	0.635	0.221	-0.03
		M	652	0.642	0.228	

Table 2
Means and Standard Deviations for Science Content Scales

Form A		Gender	N	Mean	SD	Difference in standard dev.
	Constructing	Female	686	0.524	0.205	-0.11
		Male	655	0.547	0.217	
	Reflecting	F	686	0.636	0.244	0.22
		M	655	0.580	0.269	
	Life	F	686	0.651	0.208	0.05
		M	655	0.640	0.229	
	Physical	F	686	0.486	0.202	-0.35
		M	655	0.560	0.221	
	Earth	F	686	0.451	0.208	-0.29
		M	655	0.514	0.230	
Form B						
	Constructing	F	656	0.670	0.224	-0.10
		M	621	0.694	0.235	
	Reflecting	F	656	0.509	0.245	-0.22
		M	621	0.563	0.254	
	Life	F	656	0.576	0.193	-0.12
		M	621	0.599	0.206	
	Physical	F	656	0.466	0.202	-0.17
		M	621	0.504	0.236	
	Earth	F	656	0.536	0.209	-0.23
		M	621	0.587	0.238	

To test the statistical significance of these differences, an ANOVA was conducted on the content scale scores for each subject area. Each ANOVA had one between-subjects factor, gender, one within-subject factor, scale content, and a covariate, total multiple choice score. A multivariate repeated-measures design was used because it is not based on the assumption of sphericity (equal variances of the difference scores for all pairs of levels of the repeated factor) as the univariate model is. Wilks' Λ, along with the corresponding F-approximation and probability, was reported for each effect. In general, power differences among the common multivariate test statistics (Λ, Pillais' trace, Hotelling-Lawley trace) tend to be small (Rencher, 1995; Stevens, 1992); Λ was chosen here because it also serves as a measure of effect size -- it ranges from 0, when the groups are maximally separated, to 1, when there are no differences between groups. With these large sample sizes, almost any difference would be statistically significant, so measures of the magnitude of the effects (Λ for the interaction, differences in proportions or standard deviation units for individual effects) are particularly important.

In math on form A, there was a significant content by gender interaction, though the effect was small (Λ=.993, $F_{3,1169}$=2.67, p=.0463). On form B this interaction was not significant (Λ=.998, $F_{3,1340}$=0.65, p=.5856), and the differences in the means were not in the same directions as on form A. In science the interaction was significant on form A (Λ=.989, $F_{3,1334}$=3.88, p=.0039), but not on form B (Λ=.995, $F_{3,1270}$=1.66, p=.1554), and again the relative sizes of the differences was not consistent for both forms.

High Ability Students

To learn more about a selective sample, the students whose total scores were above the 95th percentile were selected to represent high ability students. On math form A, 4.5% of the females and 5.9% of the males met this criteria; on form B, it was 3.5% of the females and 6.6%

of the males. Descriptive statistics for these groups appear in Table 18. Again, scores are given in proportions to make them easier to compare.

The differences, in standard deviation units, appear much larger with this selective group, in part because of reduced variance. In math on form A, the males scored higher on data analysis and geometry, while females scored higher on algebra and there was almost no difference on the numbers scale. Λ=.790 ($F_{3,52}$=4.61, p=.0062), a larger effect than was seen in the total group (the differences in terms of standard deviations also seemed fairly large: .81 for data analysis, .57 for geometry, -.32 for algebra). Note that males scored better on geometry (as in other studies) in this selected group, while in the total group females scored higher on geometry. On form B, males scored better on geometry and numbers, with almost no differences on algebra or data analysis (where there was a large difference on form A). The interaction between content scale and gender was not statistically significant (Λ=.934, $F_{3,63}$=1.47, p=.2302).

Table 3

Means and Standard Deviations for the Top 5% of Students on the Math Section

Form A	Gender	N	Mean	SD	Difference in standard deviations
Numbers	Females	25	0.912	0.078	0.03
	Males	31	0.910	0.079	
Data Analysis	F	25	0.864	0.104	-0.81
	M	31	0.939	0.080	
Geometry	F	25	0.892	0.104	-0.57
	M	31	0.942	0.072	
Algebra	F	25	0.952	0.051	0.32
	M	31	0.932	0.070	
Form B					
Numbers	Females	24	0.929	0.075	-0.56
	Males	43	0.965	0.057	
Data Analysis	F	24	0.904	0.075	-0.09
	M	43	0.912	0.093	
Geometry	F	24	0.942	0.078	-0.69
	M	43	0.981	0.039	
Algebra	F	24	0.950	0.066	0.02
	M	43	0.949	0.067	

In science 3.8% of the females and 6.6% of the males who took form A and 3.4% of the females and 8.1% of the males who took form B scored in the top 5%. Descriptive statistics are in Table 4.

Table 4

Means and Standard Deviations for the Top 5% of Students on the Science Section

Form A	Gender	N	Mean	SD	Difference in standard deviations
Constructing	Female	26	0.868	0.109	-0.31
	Male	43	0.902	0.103	
Reflecting	F	26	0.854	0.165	-0.19
	M	43	0.884	0.153	
Life	F	26	0.940	0.078	-0.08
	M	43	0.946	0.074	
Physical	F	26	0.831	0.112	-0.15
	M	43	0.847	0.105	
Earth	F	26	0.786	0.117	-0.34
	M	43	0.829	0.131	
Form B					
Constructing	F	22	0.924	0.093	-0.42
	M	50	0.958	0.074	
Reflecting	F	22	0.841	0.131	-0.18
	M	50	0.867	0.147	
Life	F	22	0.918	0.096	0.31
	M	50	0.884	0.113	
Physical	F	22	0.830	0.113	-0.26
	M	50	0.860	0.117	
Earth	F	22	0.854	0.099	-0.07
	M	50	0.862	0.116	

Males scored higher on every content scale on form A, with the highest difference on the Earth scale and the smallest difference on the Life scale. On form B, males again scored higher on every content scale except Life, but the *smallest* difference was on the Earth scale. However, the gender by content interaction was not statistically significant for either form (form A: Λ=.978, $F_{4,64}$=0.36, p=.8378, form B: Λ=.947, $F_{4,67}$=0.94, p=.4456).

Content Scales and Constructs Measured

Correlations

Correlation coefficients among the content scales for math form A are reported in Table 5, and the correlations for science form A are reported in Tables 6. Correlations for form B were similar.

Table 5

Correlations among Math Scales Form A

Males and Females (N = 1175)	Numbers	Data	Geometry	Algebra
Numbers	1.00000	0.64058	0.61299	0.62054
Data	0.64058	1.00000	0.60903	0.65844
Geometry	0.61299	0.60903	1.00000	0.62686
Algebra	0.62054	0.65844	0.62686	1.00000
Females (N = 603)				
	Numbers	Data	Geometry	Algebra
Numbers	1.00000	0.61874	0.56696	0.59190
Data	0.61874	1.00000	0.57533	0.62892
Geometry	0.56696	0.57533	1.00000	0.58582
Algebra	0.59190	0.62892	0.58582	1.00000
Males (N = 572)				
	Numbers	Data	Geometry	Algebra
Numbers	1.00000	0.66179	0.66046	0.64934
Data	0.66179	1.00000	0.64580	0.68881
Geometry	0.66046	0.64580	1.00000	0.67273
Algebra	0.64934	0.68881	0.67273	1.00000

Table 6

Correlations among Science Scales Form A

Males and Females (N=1347)	Constructing	Reflecting	Life	Physical	Earth
Constructing	1.00000	0.42484	0.55333	0.52284	0.54638
Reflecting	0.42484	1.00000	0.48557	0.36422	0.37821
Life	0.55333	0.48557	1.00000	0.51969	0.54300
Physical	0.52284	0.36422	0.51969	1.00000	0.51534
Earth	0.54638	0.37821	0.54300	0.51534	1.00000
Females (N=690)					
	Constructing	Reflecting	Life	Physical	Earth
Constructing	1.00000	0.36541	0.53287	0.49346	0.48713
Reflecting	0.36541	1.00000	0.43852	0.33302	0.37424
Life	0.53287	0.43852	1.00000	0.49663	0.51364
Physical	0.49346	0.33302	0.49663	1.00000	0.41690
Earth	0.48713	0.37424	0.51364	0.41690	1.00000
Males (N=657)					
	Constructing	Reflecting	Life	Physical	Earth
Constructing	1.00000	0.49590	0.57655	0.54905	0.59730
Reflecting	0.49590	1.00000	0.52745	0.44358	0.42145
Life	0.57655	0.52745	1.00000	0.56495	0.58584
Physical	0.54905	0.44358	0.56495	1.00000	0.58180
Earth	0.59730	0.42145	0.58584	0.58180	1.00000

BEST COPY AVAILABLE

In general, the correlations for each gender seemed to be quite similar to the correlations for the total group, though the correlations for males were consistently somewhat higher, especially in science. To obtain a single, composite index of the likelihood that all correlations were the same for males and females, LISREL VII was used. The fit indices are reported in Table 7.

Table 7

Fit Indices for Model of Equal Correlation Matrices for Males and Females

		GFI female	GFI male	RMSR female	RMSR male	χ^2	df	prob
Math								
	Form A	.994	.991	.027	.028	17.82	14	.215
	Form B	.997	.996	.014	.015	9.82	14	.775
Science								
		.990	.986	.039	.040	39.75	20	.005
		.991	.987	.036	.038	34.76	20	.021

All of the Goodness of Fit Indices (GFI) were greater than .98, indicating that the model with equal correlation matrices was quite tenable. The χ^2 values were significant (suggesting poor fit) on the science test, but with this large sample size it took only small differences to reach statistical significance. The root mean square residual was lower on the math tests than on the science tests, but seemed reasonably small for both subject areas.

Standard errors, based on Cronbach's alpha, were also estimated for each content scale; they are reported here relative to percentage scores, not raw scores. Differences in the standard error of measurement could indicate more random variance was affecting one gender. These standard errors and the corresponding reliabilities are reported in Tables 8 and 9.

In math, the standard errors seemed about the same for males and females (Table 8); the reliabilities tended to be somewhat higher for males (with the exception of algebra on form A and algebra for the subgroup of responders on form B). The same pattern was followed in science

(Table 9), where the differences in reliabilities were greater (the males had greater variances, so with comparable standard errors the reliabilities were higher). Also, on both forms the standard error of measurement for the physical science scale was greater for females.

Table 8

Standard Error of Measurement and Reliability for Math

	# of items on scale	males and females		females		males	
		std. error	reliability	std. error	reliability	std. error	reliability
Form A							
Numbers	10	0.134	.618	0.133	.601	0.134	.636
Data	10	0.135	.645	0.134	.640	0.135	.652
Geometry	10	0.133	.595	0.133	.559	0.132	.627
Algebra	10	0.133	.694	0.133	.696	0.133	.692
Form B							
Numbers	10	0.131	.698	0.131	.672	0.131	.724
Data	10	0.136	.629	0.136	.610	0.135	.649
Geometry	10	0.133	.717	0.134	.695	0.132	.736
Algebra	10	0.136	.631	0.137	.618	0.136	.646

Table 9

Standard Error of Measurement and Reliability for Science

	# of items on scale	males and females		females		males	
		std. error	reliability	std. error	reliability	std. error	reliability
Form A							
Constructing	9	0.142	.550	0.141	.528	0.142	.571
Reflecting	5	0.201	.393	0.200	.328	0.202	.438
Life	9	0.140	.588	0.140	.544	0.140	.627
Physical	10	0.145	.525	0.146	.481	0.143	.542
Earth	9	0.147	.561	0.148	.493	0.145	.605
Form B							
Constructing	9	0.136	.646	0.138	.619 --	0.134	.674
Reflecting	6	0.178	.496	0.178	.470	0.176	.522
Life	10	0.134	.549	0.135	.513	0.133	.582
Physical	8	0.158	.480	0.161	.365	0.155	.566
Earth	9	0.153	.539	0.155	.451	0.150	.605

Summary and Conclusions

On the math tests the gender by content interaction was significant on one of the two forms, and the differences were small on both forms. Females scored higher on geometry on form A (and lower on everything else), but on form B the males scored higher on every scale, with the largest difference on geometry and the smallest on algebra. The findings on Form A were opposite what would be expected from other studies. The findings from form B, of greater differences on geometry than algebra, are consistent with findings for high school students on the SAT and ACT (Doolittle & Cleary, 1987; Harris & Carlton, 1993), but the differences were small and inconsistent in this study. The students taking the HSPT were from a broader ability range than the students taking the SAT or ACT (especially Doolittle and Cleary's sample of students who had completed a precalculus or trigonometry course). The content of the HSPT is also at a somewhat lower level.

In science, the gender by content interaction was statistically significant on form A. However, the patterns of differences were inconsistent across forms. On form A females scored higher on reflecting on knowledge and life science, while males scored higher on the other scales, particularly physical and earth sciences. This is fairly consistent with other research, where the smallest differences, or differences in favor of females, tend to be on life science and process-type scales or tests (Becker, 1989; Comber & Keeves, 1973; Erickson & Erickson, 1984; Linn, De Benedictis, Delucchi, Harris, & Stage, 1987; Linn & Hyde, 1989; Stanley, 1987), but again the sizes of the differences in this study were small. On form B, in contrast, the male advantage on reflecting was as high as it was on the physical and earth scales. Several of the studies cited above (Comber & Keeves, 1973; Erickson & Erickson, 1984; Linn, De Benedictis, Delucchi, Harris, & Stage, 1987; Linn & Hyde, 1989) used samples of students of all abilities and tested

content appropriate for typical high school students. There must be something else unique about the HSPT (or at least this form of the HSPT) which produced a different pattern.

One possible reason for the inconsistent findings in this study might be that individual items which showed gender differences not predicted from the total score distribution (as well as items judged to appear biased) were detected and modified or eliminated through earlier tryouts. Therefore, the remaining gender differences tended to be small (after total score was controlled) and their fluctuations across forms would be due to chance. Note that the total score, not individual scale scores, was used as the basis for DIF analysis. If the subscores for each scale had been used to identify items which showed DIF, fewer items might have been identified and greater gender differences between scales might have been observed.

Looking only at the students in the top 5% of the sample, in math on both forms males scored higher on the geometry scale and either there were no differences on algebra or females scored higher (and the content by gender interaction was statistically significant for one form). This pattern was consistent with findings for high school students on the SAT and ACT (Doolitle & Cleary, 1987; Harris & Carlton, 1993). However, differences on the other two scales were not consistent across forms. This illustrates the importance of looking at more than one form (assuming generalizations are to be made to a class of items) before drawing general conclusions. In science, the pattern of differences on the content scales was *not* consistent across the two forms, except that females did relatively better on the Life scale. No general conclusions about the gender by content interaction can be made at this point.

The correlations among the content scales did not vary appreciably by gender, and the standard errors were similar across gender.

The major finding of the study was that there do not seem to be consistent gender by content differences, when ability (represented by total test score) is controlled.

References

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two meta-analyses. Journal of Research in Science Teaching, 26, 141-169.

Becker, B. J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. American Educational Research Journal, 27, 65-87.

Comber, L. C., & Keeves, J. P. (1973). Science education in 19 countries: An empirical study. New York: John Wiley & Sons.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24, 157-166.

Erickson, G. L., & Erickson, L. J. (1984). Females and science achievement: Evidence, explanations, and implications. Science Education, 68, 63-89.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. Review of Educational Research, 62, 61-84.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. Applied Measurement in Education, 6, 137-151.

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. Review of Educational Research, 63 (1), 94-105.

Katzman, S., & Alliger, G. M. (1992). Averaging untransformed variance ratios can be misleading: A comment on Feingold. Review of Educational Research, 62, 427-428.

Linn, M. C., De Benedictis, T., Delucchi, K., Harris, A. & Stage, E. (1987). Gender differences in National Assessment of Educational Progress items: What does "I don't know" really mean? Journal of Research in Science Teaching, 24, 267-278.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. Educational Researcher, 18, 17-19, 22-27.

Ryan, K. E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. Educational Measurement: Issues and Practice, 15(4), 15-20, 38.

Stanley, J. C. (1987, April). Gender differences on the College Board Achievement Tests and the Advanced Placement Examinations: Effect sizes versus some upper-tail ratios. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 286 913)